TopicWeave: Enhancing Scientific Literature Topic Modeling through Distilled Model Ensembles - Final Report

Quddus Chong
MIDS, UC Berkeley
quddus.chong@berkeley.edu

Instructor: Jennifer Zhu
MIDS, UC Berkeley
jennifer.zhu@berkeley.edu

Abstract

We introduce TopicWeave, a novel approach for scientific literature topic modeling that employs a distilled model ensemble of domain-specific (SciBERT) and citation-aware (SPECTER) representations, further enhanced by domain-adaptation of SciBERT. By combining these complementary embeddings, fine-tuning on a scientific corpus, and optimizing hyperparameters, Topic Weave significantly outperforms a strong baseline on a 50,000-paper arXiv dataset, achieving a 36% relative improvement in Normalized Mutual Information and a 489% relative improvement in Adjusted Rand Index. This results in more coherent, interpretable, and scientifically grounded topics, demonstrating the potential of our ensemble approach to advance the state of the art in scientific text analysis.

1 Introduction

The rapid expansion of scientific literature necessitates efficient methods for organization and insight extraction. Topic modeling offers a valuable approach to uncover latent thematic structures, aiding in tracking trends and reviewing literature. These structures also enhance information extraction and named entity recognition, improving the ability to identify entities and relationships within scientific text. Neural topic modeling approaches like BERTopic have shown effectiveness in general applications but often fail to capture domain-specific nuances in scientific literature due to its specialized vocabulary and frameworks. TopicWeave addresses this by proposing an ensemble approach that combines distilled versions of SciBERT and SPECTER to create a unified input representation for BERTopic. By leveraging both SciBERT's domain-specific understanding and SPECTER's citation-based relational knowledge, TopicWeave aims to produce more coherent, interpretable, and scientifically relevant topics while maintaining computational efficiency through model distillation. We further enhance performance through domain adaptation fine-tuning and hyperparameter optimization.

2 Background

Neural topic modeling has evolved significantly from traditional probabilistic methods like Latent Dirichlet Allocation (LDA). (Abdelrazek et al., 2023) provide a comprehensive survey of topic modeling algorithms and applications, highlighting the shift toward neural approaches. BERTopic (Grootendorst, 2022) represents a significant advancement by leveraging neural embeddings to cluster documents and extract topics using TF-IDF, demonstrating superior performance compared to traditional methods.

Recent work has recognized the value of domain-specific adaptations for scientific literature. (Chagnon et al., 2024) introduced BERTeley, which streamlines topic modeling for scientific literature. However, as noted by (Kang et al., 2023), there remains significant potential for improving topic modeling through specialized language models. (Wolff et al., 2024) demonstrated that enriched BERT embeddings can enhance scholarly publication classification, providing evidence that domain-specific pre-training yields benefits for scientific document understanding.

The potential of ensemble approaches has been explored by (Voskergian et al., 2024), who showed that combining multiple topic models can improve text classification performance. However, their approach focuses on post-hoc ensemble methods rather than the integration of complementary embedding models at the representation level as proposed in TopicWeave.

Distillation techniques offer a promising avenue for maintaining the benefits of large pre-trained models while reducing computational requirements. The SciBERT model (Beltagy et al.,

2019) was specifically pre-trained on a large corpus of scientific text and has demonstrated superior performance on domain-specific tasks. Similarly, SPECTER2 (Singh et al., 2022) captures inter-document relationships through training on citation graphs, providing complementary information about document relationships that pure text-based models may miss.

Recent work by (Mu et al., 2024) suggests that large language models can offer alternative approaches to traditional topic modeling, indicating a broader trend toward using advanced language models for this task. (Wijanto et al., 2024) explored optimal hyperparameter tuning for BERT-based topic models applied to scientific articles, highlighting the importance of careful parameter selection for these approaches.

TopicWeave builds on these advances by combining domain-specific embeddings through an ensemble approach while ensuring computational efficiency through model distillation. The key innovation in our neural architecture is the knowledge-specialized dual pathway approach, where two parallel embedding models capture different aspects of scientific understanding: domain-specific vocabulary (SciBERT) and document relationships (SPECTER).

3 Methods

3.1 Dataset Acquisition and Preprocessing

We extracted a balanced subset of 50,000 articles in 10 scientific categories from the arXiv data set (Cornell University, 2024), obtained from Kaggle. From this raw dataset, we implemented an independent preprocessing pipeline that retained document IDs, titles, abstracts, and primary categories. This includes text cleaning with stopword removal and lemmatization. The titles and abstracts were concatenated to create single document representations, with titles repeated twice to emphasize key terminology. The resulting document representations were split into training (90%) and validation (10%) sets. Author-assigned categories served as external validation metrics.

3.2 Baseline Model Implementation

A baseline BERTopic model was implemented using the *paraphrase-MiniLM-L12-v2* sentence transformer as the embedding model, without domain-specific pretraining, configured with a minimum topic size of 10 documents to ensure meaningful topic representations. The baseline followed BERTopic's standard pipeline: embedding generation, dimensionality reduction using UMAP, clustering with HDBSCAN, and topic extraction using c-TF-IDF.

3.3 TopicWeave Architecture Implementation

The TopicWeave architecture involves a multi-stage approach, illustrated by Figure 1:

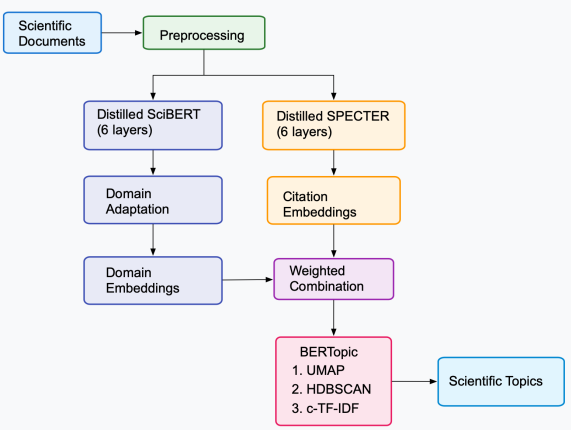

Figure 1: TopicWeave architecture showing the dual pathway approach combining domain-specific and citation-aware embeddings.

1. **Model Distillation**: We created two distilled domain-specific models:

- **Distilled SciBERT**: Reduced from 12 to 6 transformer layers using the original vocabulary but training new embeddings. The layer reduction aimed to maintain 95% of the full model's performance while halving computational costs.

- **Citation-aware Model**: Initially, a 6-layer contrastive learning model (Gao et al., 2021) was developed for semantic relationships but underperformed due to lack of citation context. We then implemented a 6-layer distilled version of SPECTER (Cohan et al., 2020), pre-trained on citation graphs for inter-paper relationships, which better captured citation-grounded semantics. The original SPECTER was chosen over SPECTER2 due to PEFT framework compatibility issues.

2. **Embedding Generation**: Each model independently generates embeddings for all preprocessed documents, capturing complementary aspects of scientific text.

3. **Optimized Ensemble Weighting**: A weighted average was used to combine SciBERT and SPECTER embeddings into a single unified representation. The weights were optimized to maximize a combined score that balanced several evaluation criteria. This combined score was calculated as: 0.3 * silhouette score + 0.3 * NMI + 0.2 * ARI + 0.2 * within-category similarity. Silhouette score was calculated using cosine similarity on a random sample of documents, while NMI and ARI were computed after encoding document categories into numerical labels and performing KMeans clustering. Within-category similarity measured the average cosine similarity of embeddings within the same category, also calculated on a random sample. The optimization process involved iterating through SciBERT weights from 0.0 to 1.0, with the SPECTER weight being the complement (1.0 - SciBERT weight), and selecting the weight that yielded the highest combined score.

4. **Domain Adaptation**: We further enhanced the distilled SciBERT model through domain-specific fine-tuning. We employed masked language modeling, a technique where the model is trained to predict randomly hidden words, to adapt it to our corpus of 50,000 arXiv papers. During fine-tuning, 15% of tokens in the input sequences were masked. The model was trained for 3 epochs with the following hyperparameters: learning rate = 5e-5, batch size = 8, AdamW optimizer, and weight decay = 0.01. This fine-tuning on scientific abstracts enables the model to better capture general scientific discourse patterns, domain-specific terminology, and subtle contextual relationships within scientific text.

5. **BERTopic Integration**: The combined embeddings were integrated with BERTopic's standard pipeline: dimensionality reduction using UMAP, clustering with HDBSCAN, and topic extraction using c-TF-IDF.

Distillation reduced model parameters by 50%, resulting in an estimated 30-40% speedup in embedding generation time and 30-50% reduction in memory requirements. This computational efficiency is critical for practical applications involving large scientific corpora.

We implemented this pipeline using Python and PyTorch, with the Hugging Face Transformers and SentenceTransformer libraries. All experiments were run on Google Colab with A100 GPUs.

3.4 Evaluation Metrics

We evaluated models using both quantitative alignment metrics (NMI, ARI) and qualitative coherence metrics (C_v, NPMI):

Normalized Mutual Information (NMI) (McDaid et al., 2011) serves as our primary evaluation metric, measuring the normalized mutual information (scaled to [0,1]) between discovered topic clusters and ground truth categories, where 0 indicates no relationship and 1 perfect alignment. We prioritize NMI due to its effectiveness in handling varying cluster numbers, insensitivity to specific label assignments, and ability to capture partial matches between topics and categories, which is crucial in scientific literature with fluid disciplinary boundaries. Higher NMI signifies better topic-to-domain correspondence.

Adjusted Rand Index (ARI) (Santos and Embrechts, 2009) complements NMI by quantifying the similarity of document pair assignments; it counts how often document pairs are placed in the same topic and category versus different ones. Ranging from -1 to 1 (1: perfect agreement, 0: random, -1 to 0: worse than random), ARI emphasizes exact matching of clustering assignments. While strict about near-misses, ARI helps assess the precision of topic boundary alignment with established categories.

Topic Coherence (C_v and NPMI) (Röder et al., 2015) assesses the semantic interpretability of topics based on word co-occurrence patterns. C_v combines indirect cosine similarity and NPMI to measure this. NPMI (Normalized Pointwise Mutual Information) specifically measures the statistical co-occurrence of words, normalized to [-1,1]. These coherence metrics correlate with human judgment of topic quality and capture semantic relationships that clustering metrics might overlook.

Together, these metrics form a balanced evaluation framework, addressing both structural validity (alignment with categories) and semantic quality (topic coherence). By prioritizing NMI while monitoring coherence metrics, we balance the dual objectives of maintaining alignment with established scientific classifications while discovering meaningful semantic patterns within the literature.

3.5 Experiments and Fine-tuning

We conducted extensive hyperparameter tuning for the BERTopic model used to analyze embeddings from domain-adapted SciBERT and Top-

icWeave. We optimized the minimum topic size (5-15), UMAP dimensionality (5-10), UMAP neighbor count (10-20), and HDBSCAN minimum samples (5-15). Best parameters were determined by maximizing a combined score based on NMI, ARI, and coherence metrics. Sensitivity analysis revealed that UMAP parameters had the greatest impact on performance, with diminishing returns beyond 10 components. These optimal parameters were applied across all scientific domains, and the results suggest good generalizability of our approach across different scientific fields.

4 Results and Discussion

4.1 Model Performance Comparsion

The results in Table 1 demonstrate that TopicWeave consistently outperforms baseline approaches across all metrics. Our Domain-Adapted TopicWeave (Tuned) model achieves substantial improvements with 36% higher NMI (0.411 vs 0.303) and 489% higher ARI (0.165 vs 0.028) compared to the baseline, indicating a much stronger alignment with ground truth categories. Furthermore, TopicWeave produced more coherent topics (NPMI: +24%, C_v: +6%).

The component models exhibit complementary strengths: SciBERT produces fewer topics (530) with high coherence (C_v: 0.620) but limited category alignment (NMI: 0.162), while the SPECTER model yields many topics (380) with stronger category correlation (NMI: 0.335).

TopicWeave effectively combines these strengths, with hyperparameter tuning further refining topic identification to 247 topics, avoiding the baseline's issue of a large outlier topic. Notably, domain adaptation significantly enhances TopicWeave's performance, with the tuned TopicWeave model achieving substantial gains in both coherence (C_v: +4%, NPMI: +10%) and category alignment (NMI: +6%, ARI: +97%) compared to the original TopicWeave model.

4.2 Topic Quality Analysis

Our qualitative analysis reveals significant differences in topic quality across models. The baseline model suffered from extreme topic imbalance with its largest topic containing 11,068 documents (49% of corpus), while TopicWeave's largest topic includes only 4,559 documents (10%), indicating more effective document differentiation.

Table 2 summarizes the top topics discovered by the Domain-Adapted TopicWeave (Tuned) model.

TopicWeave's top topics demonstrate genuine scientific discourse patterns through domain-specific terminology rather than generic terms. The topics reveal specialized research approaches (numerical simulations, Bayesian inference, kernel classification) and methodological connections across disciplines, capturing both field-specific concepts and cross-cutting analytical techniques used throughout scientific literature.

Figure 2: Topic-category alignment, revealing both category-specific topics and interdisciplinary themes spanning multiple scientific domains.

Figure 2 shows TopicWeave's topic distribution across arXiv categories, demonstrating its nuanced understanding of scientific literature. While Topic 0 (numerical simulations) shows a strong concentration (e.g., ∼24% within physics.comp-ph), indicating domain relevance, TopicWeave also identifies significant interdisciplinary links. For instance, Topic 5 (reinforcement learning) exhibits notable proportions across cs.LG, cs.AI, cs.CL, cs.RO, and physics.comp-ph, highlighting its broad applicability. Furthermore, Topic 4 (statistical processes) bridges computational linguistics (cs.CL, ∼30%) and quantitative biology (q-bio.QM, significant proportion), revealing shared methodological foundations.

5 Conclusion

Our work demonstrates that TopicWeave's ensemble approach with distilled domain-specific models provides meaningful improvements in scientific literature topic modeling while maintaining

Table 1: Performance Comparison of Topic Models

Model	NMI	ARI	Topics	C_v Coherence	NPMI Coherence
Base Models					
Baseline	0.303	0.028	346	0.580	0.098
TopicWeave (Original)	0.335	0.031	370	0.600	0.103
SciBERT (Original)	0.162	0.011	530	0.620	0.110
SPECTER	0.335	0.037	380	0.588	0.099
Domain-Adapted Models					
Domain-Adapted SciBERT	0.316	0.028	315	0.592	0.110
Domain-Adapted TopicWeave	0.358	0.067	301	0.604	0.114
Hyperparameter-Tuned Models					
Domain-Adapted SciBERT (Tuned)	0.397	0.127	165	0.640	0.142
Domain-Adapted TopicWeave (Tuned)	**0.411**	**0.165**	247	0.616	0.122
Improvement Over Baseline					
Absolute Improvement	+0.108	+0.137	-99	+0.036	+0.024
Percentage Improvement	+36%	+489%	-29%	+6%	+24%

Table 2: Top 10 Topics - Domain-Adapted TopicWeave (Tuned). Primary Category indicates the arXiv category where each topic has its highest concentration (percentage of topic documents in that category). Prevalence shows the topic's overall representation in the corpus by percentage and document count.

Topic	Representative Terms	Primary Category	Prevalence
0	numerical, simulations, equation, equations, energy	physics.comp-ph (24%)	10.0% (4559 docs)
1	research, citation, scientific, science, papers	q-bio.QM (18%)	7.8% (3571 docs)
2	market, stock, financial, price, markets	q-fin.ST (18%)	4.6% (2093 docs)
3	classification, kernel, classifier, classifiers, learning	cs.AI (28%)	3.1% (1406 docs)
4	processes, asymptotic, process, stationary, series	cs.CL (30%)	1.0% (450 docs)
5	reinforcement, policy, learning, exploration, agent	cs.LG (15%)	0.9% (399 docs)
6	causal, graphical, independence, variables, acyclic	math.ST (16%)	0.8% (387 docs)
7	sentiment, tweets, twitter, polarity, social	cs.LG (18%)	0.8% (377 docs)
8	speech, recognition, acoustic, speaker, asr	cs.LG (17%)	0.8% (352 docs)
9	belief, bayesian, inference, propagation, networks	q-bio.QM (20%)	0.7% (341 docs)

computational efficiency. Our results show significant improvements over baseline methods, with the Domain-Adapted TopicWeave (Tuned) model achieving 36% higher NMI and 489% higher ARI scores alongside improved topic coherence.

Future work could explore more sophisticated approaches to citation modeling using Graph Neural Networks, learnable embedding combinations through attention mechanisms, and hierarchical topic structures to better represent the organization of scientific knowledge. (Belfathi et al., 2024) highlights the need for domain-specific adaptations of language models. This points to potential future improvements to TopicWeave by incorporating selective masking strategies. (Rahimi et al., 2023) proposes the Contextualized Topic Coherence approach to use large language models for evaluating topic coherence, which could lead to a more accurate and human-aligned evaluation of topic coherence for TopicWeave.

References

Aly Abdelrazek et al. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Anas Belfathi et al. 2024. Language model adaptation to specialized domains through selective masking based on genre and topical characteristics. *arXiv preprint arXiv:2402.12036*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Eric Chagnon et al. 2024. Benchmarking topic models on scientific articles using berteley. *Natural Language Processing Journal*, 6:100044.

Arman Cohan, Waleed Naseh, Hamed Zamani, and W Bruce Cro Croft. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7201–7208.

Cornell University. 2024. arxiv. Kaggle, 27 Feb. 2024, www.kaggle.com/datasets/Cornell-University/arxiv.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Woojin Kang et al. 2023. An analysis of research trends on language model using bertopic. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*. IEEE.

Aaron F McDaid, Derek Greene, and Neil Hurley. 2011. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*.

Yida Mu et al. 2024. Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248*.

Hamed Rahimi et al. 2023. Contextualized topic coherence metrics. *arXiv preprint arXiv:2305.14587*.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*.

Jorge M Santos and Mark Embrechts. 2009. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International conference on artificial neural networks*, pages 175–184, Berlin, Heidelberg. Springer Berlin Heidelberg.

Amanpreet Singh et al. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

Daniel Voskergian, Rashid Jayousi, and Malik Yousef. 2024. Topic selection for text classification using ensemble topic modeling with grouping, scoring, and modeling approach. *Scientific Reports*, 14(1):23516.

Maresha Caroline Wijanto, Ika Widiastuti, and Hwan-Seung Yong. 2024. Topic modeling for scientific articles: Exploring optimal hyperparameter tuning in bert. *International Journal on Advanced Science, Engineering & Information Technology*, 14(3).

Benjamin Wolff, Eva Seidlmayer, and Konrad U Förstner. 2024. Enriched bert embeddings for scholarly publication classification. In *International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs*. Springer Nature Switzerland.